RAMS
Reliability, Availability,
Maintainability, and Safety

Compering the ACER and POT MCMC (acronyms?) Extreme Value Statistics Methods Through Analysis of Commodities Data

Kristoffer Kofoed Rødvei

April 2016

MASTER THESIS

Department of Production and Quality Engineering (What here??)
Norwegian University of Science and Technology

Supervisor: Professor Arivd Næss

Secondary Advisor: Professor Andre Riebler

Helping(???): Professor Sjur Westgaard

Preface

Here, you give a brief introduction to your work. What it is (e.g., a Master's thesis in RAMS at NTNU as part of the study program xxx and...), when it was carried out (e.g., during the autumn semester of 2021). If the project has been carried out for a company, you should mention this and also describe the cooperation with the company. You may also describe how the idea to the project was brought up.

It is suggested that the reader have some statistical knowledge, especially within extreme value and computational statistical. Some financial knowledge and experience with value at risk is desirable.

"When your done: https://www.scribbr.com/thesis/preface-thesis/"

Trondheim, 2016-04-25

(Your signature)

Kristoffer Kofoed Rødvei

Summary and Conclusions

Here you give a summary of your work and your results. This is like a management summary and should be written in a clear and easy language, without many difficult terms and without abbreviations. Everything you present here must be treated in more detail in the main report. You should not give any references to the report in the summary – just explain what you have done and what you have found out. The Summary and Conclusions should be no more than two pages.

You may assume that you have got three minutes to present to the Rector of NTNU what you have done and what you have found out as part of your thesis. (He is an intelligent person, but does not know much about your field of expertise.)

Contents

Preface . . . i

Summary and Conclusions . . . ii

1 Theory 2

1.1 Extreme value theory . . . 2

1.1.1 Peak Over Threshold . . . 4

1.1.2 Average Condition Exceedance Rate . . . 7

1.1.3 Predicting Future Extreme With Confidence Interval . . . 10

1.2 Bayesian Inference . . . 10

1.3 Markov Chain Monte Carlo Method . . . 13

1.3.1 Gibbs Sampling . . . 13

1.3.2 Metropolis–Hastings Algorithm . . . 14

1.3.3 Effective Sample Size . . . 15

1.3.4 Adaptive Metropolis Algorithm . . . 15

1.3.5 Applying Markov Chain Monte Carlo to Peak Over Threshold . . . 17

1.3.6 Multivariate Random Normal Generator . . . 19

1.4 Value at Risk . . . 19

1.5 Evaluating Forecasts . . . 19

1.6 ARCH/GARCH . . . 20

2 Data 21

2.1 Synthetic Data . . . 21

2.1.1 Pareto . . . 21

2.1.2 Commodity Generated Data . . . 21

2.2 Commodity Data 21

3 Summary **22**

3.1 Summary and Conclusions 22

3.2 Discussion 23

3.3 Recommendations for Further Work 23

A Additional Information **24**

A.1 Introduction 24

A.1.1 R 24

A.1.2 C++ 24

Bibliography **26**

Chapter 1

Theory

The following chapter gives an introduction to the theory behind this work. Some of the sections only give a brief description of the theory with sources for more in depth explanation.

1.1 Extreme value theory

The basics of extreme value theory, is to analyze the maximum of a series of random variables $X_1, \dots, X_n$. Defining M_n as the maximum of a series

$$M_n = \max\{X_1, \dots, X_n\}, \tag{1.1}$$

the resulting distribution of M_n is

$$\Pr(M_n \leq z) = \Pr(X_1 \leq z, \dots, X_n \leq z). \tag{1.2}$$

By assuming $X_1, \dots, X_n$ independent and identically distributed (i.i.d.) with common cumulative distribution function F, equation (1.2) reduces to

$$\begin{aligned} \Pr(M_n \leq z) &= \Pr(X_1 \leq z) \times \cdots \times \Pr(X_n \leq z) \\ &= \left[F(z)\right]^n. \end{aligned} \tag{1.3}$$

The distributing F is normally unknown, and a divergent in the estimated distributing F could potentially escalate to a large difference in the resulting F^n.

As for the central limit theory for the normal distribution, F^n also has a limiting distribution, by The Fisher–Tippett–Gnedenko theorem. If there exist an $a_n > 0$ and b_n such that

$$\lim_{n\to\infty} \Pr\left[(M_n - b_n)/a_n \leq z\right] \to G(z), \tag{1.4}$$

where G is a non-degenerating function, then G follows either

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \qquad -\infty < z < \infty; \tag{1.5}$$

$$G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases} \tag{1.6}$$

$$G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)\right]^{\alpha}\right\}, & z \leq b, \\ 1, & z > b; \end{cases} \tag{1.7}$$

for each case, $\alpha > 0$. Here equation (1.5), (1.6) and (1.7) refears to Gumbel, Fréchet and Weibull distribution respectively. The above equation can be combined into the General Extreme Value (GEV) distribution

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{1.8}$$

where $1 + \xi(z-\mu)/\sigma > 0$, the location parameter is $-\infty < \mu < \infty$, the scale parameter is $\sigma > 0$ and the shape parameter is $-\infty < \xi < \infty$. It can easily be verified that the GEV equals equation (1.6) when $\xi > 0$, equation(1.7) when $\xi < 0$ and converge towards equation(1.5) as $z \to 0$.

The GEV model requires i.i.d. data points for parameter estimations. Unfortunately, in practice that is rarely the case. Block maxima or r largest order statistics are common methods for filtering the dependent data points into an approximate i.i.d. dataset (Coles, 2001, p. 66). The basic principle is to only use the largest, or r largest data within each block. Examples of block sizes could be week, month, year etc.

For a deeper description of extreme value theory, GEV or block maxima, the book of (Coles, 2001, Chapter 3) is suggested.

1.1.1 Peak Over Threshold

One of the problem with the GEV method and the i.i.d. filtration of data points, is that the block maxima and r largest order statistics are quite wasteful. Specially in situations where some blocks contains more extreme than others, large extreme will be discarded from the set, which else would have been accepted in other blocks.

The Peak Over Threshold (POT) method is suggesting a different method of tackling the i.i.d. filtration. Filtering the data, by only using points over a certain threshold u, avoid the problem of discarding large extremes in certain blocks. As long as u is chosen sufficiently large, the resulting data will be i.i.d. Using the GEV distribution equation (1.8) it can be shown, like was done by (Coles, 2001, p. 76), that

$$\Pr(X > y+u | X > u) = \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{1.9}$$

where y is the threshold excess, given by $y = z - u$ for $z > u$. The resulting distribution of y is called the generalized Pareto distribution (GPD)

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{1.10}$$

or

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right) \tag{1.11}$$

when $\xi \to 0$. Here ξ equals the GEV parameter, while $\tilde{\sigma} = \sigma + \xi(u - \mu)$. The conditional probability

$$\begin{aligned}\Pr(X > y+u | X > u) &= \frac{\Pr(X > y+u, X > u)}{\Pr(X > u)} \\ &= \frac{\Pr(X > y+u)}{\Pr(X > u)},\end{aligned} \tag{1.12}$$

since $y + u \geq u$, the probability $\Pr(X > y+u, X > u) = \Pr(X > y+u)$. By combining equation

(1.9) and (1.12), the probability of a future event can be found by

$$\begin{aligned}\Pr(X > z) &= \Pr(X > y + u)\\ &= \Pr(X > u)\cdot P(X > y + u | X > u)\\ &= \Pr(X > u)\cdot\left[1 + \xi\left(\frac{z-u}{\tilde{\sigma}}\right)\right]^{-\frac{1}{\xi}},\end{aligned} \tag{1.13}$$

where $\Pr(X > u)$ is the probability that a random point exceeds the threshold.

For more information about the POT method, see (Coles, 2001, Chapter 4).

ADD/TALK/NAME GENERALIZED PARETO DISTRIBUTION (GPD) (H(y)).

y=x-u CALLED THRESHOLD EXCESS. INCLUDE IN OTHER PARTS OF THE THESIS!

Declustering

The numbers of threshold excess y increase as the threshold u decrease. A larger number of threshold excess will increase the accuracy and lower the variance of the parameter estimation, which suggest using a low threshold. In practice, data are often correlated, heteroscedastic or nonstationary. For data without trend, a high enough threshold will ensure close to i.i.d. property for the threshold excess. As the threshold decrease, clusters could appear, and the threshold excess will no longer be i.i.d. Violation of the i.i.d. property will result in an estimation bias, which suggest using a high threshold. The selection of threshold comes down to the trade-off between accuracy and bias. The goal is to get the lowest variance without bias.

Declustering method can be applied to improve the i.i.d. property for low threshold. The target is to localize clusters above the threshold and select the largest value within each cluster. The method used here, defines a cluster as the points above the threshold, until r consecutive points are observed below. Referring to figure 1.1, for an example on how the method is used in practice. For $r = 1$ there are 7 clusters, while for $r = 4$ there are 3. For the particular threshold used in the plot, $r = 4$ seems like the superior choice.

For a more in depth description of decluttering, see (Coles, 2001, p. 100).

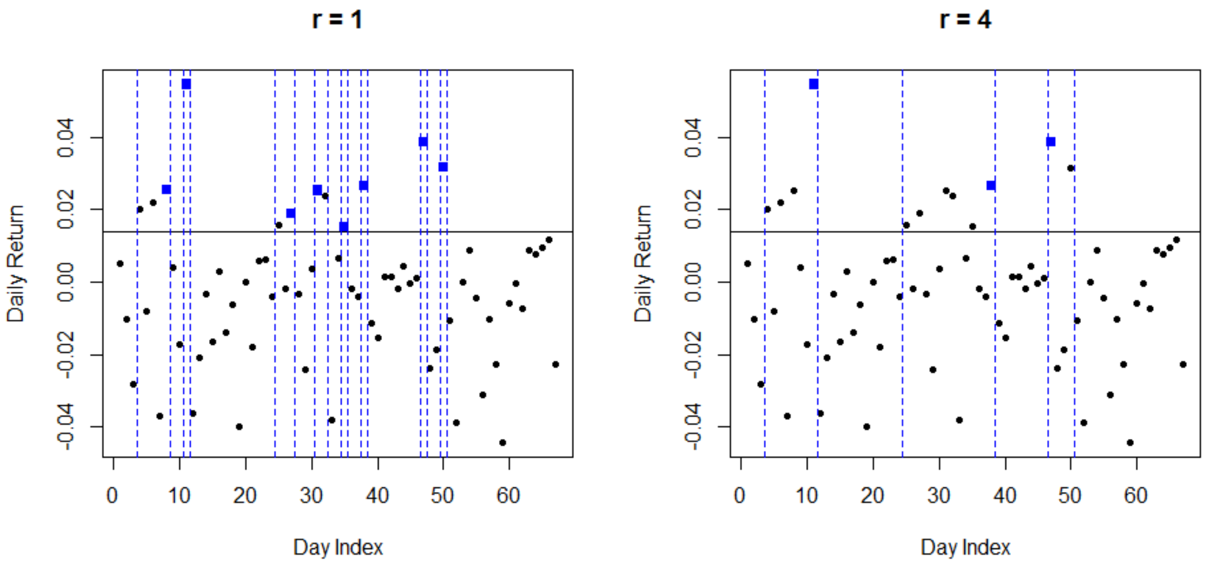

Figure 1.1: Portion of the crude oil daily return series, described in chapter 2.2. The horizontal solid line is threshold, with $u = 0.014$. Clusters are localized between the vertical blue dashed lines, with the largest value within each clusters shown as a blue square.

Threshold

As stated above, the goal when selecting threshold u is to find the smallest threshold u_0, for which the model is still unbiased, such that the highest accuracy is achieved. For this paper, the combination of two methods are used in the selection process.

The first method uses the fact that the mean of the GPD equals

$$E(Y) = \frac{\tilde{\sigma}}{1-\xi}, \tag{1.14}$$

for $\xi < 1$, and infinite when $\xi > 1$. Thus the first method fails when $\xi > 1$, but in practice ξ rarely exceeds 1. As shown above, the GPD ξ equals the GEV parameter which is independent of threshold, while $\tilde{\sigma} = \sigma + \xi(u - \mu)$ is linear with respect to threshold. Here σ and μ are the GEV parameters and independent of threshold. Thereby the mean of Y is also linear proportional to threshold. By plotting the mean of threshold excess against thresholds, linear effect should be apparent from u_0. Confidence interval can be added for a better understanding of where the linearity starts. For larger values of thresholds, there will only be a few numbers of threshold excess, hence it is suggested using t-distribution for more realistic confidence intervals. REFERING TO

PLOT BLABLABLA. For more information about the method, see (Coles, 2001, p. 79).

For the second method, estimates of ξ and $\tilde{\sigma}$ is taken for a variety of thresholds. The parameters ξ and the reparametrized $\sigma^* = \tilde{\sigma} - \xi u$ should both be constant from u_0. For pinpointing of u_0, both ξ and σ^* is plotted against u, with added confidence intervals. REFERING TO PLOT BLABLABLA. For more information about the method, see (Coles, 2001, p. 83).

After threshold selection, $\tilde{\sigma}$ is simply estimated from the threshold excess and is required larger than zero. For simplicity, from here, the notation σ is used for the GPD parameter $\tilde{\sigma}$, as long as otherwise is not specified.

1.1.2 Average Condition Exceedance Rate

The Average Conditional Exceedance Rate (ACER) is another extreme value method, first introduced by Næss and Gaidai (2009), see the paper and Næss et al. (2013) for a more in depth description of the theory, as the following is only a brief introduction. Both the GEV and GPD distributions requires the observations to be i.i.d. When observations are not i.i.d., filtering methods such as, threshold exceedance, declustering, blocking, etc., is used to achieve close to i.i.d. data. The problem with these filtering methods, is that they often discard most of the data, such that only a small amount of the data can be used for parameter estimation. The advantage of the ACER method is that the observation is not restricted to i.i.d. or even stationarity data, only requires no trend. Another advantage is the ACER methods ability to a certain extent capture the subasymtotic parts, which also could improve estimation.

Without the i.i.d. assumption for $X_1, \ldots, X_n$, equation (1.2) can be written using time dependency

$$\Pr(M_n \leq z) = \prod_{j=1}^{n} \Pr(X_j \leq z, |X_{j-1} \leq z, \ldots, X_1 \leq z) \cdot \Pr(X_1 \leq z). \tag{1.15}$$

It is reasonable to assume that the data dependency with neighboring points degrease by time, and is negligible after $k \ll n$ steps, such that

$\Pr(X_j \leq z, |X_{j-1} \leq z, \ldots, X_1 \leq z) \approx \Pr(X_j \leq z, |X_{j-1} \leq z, \ldots, X_{j-k+1} \leq z)$ for every $j = k, \ldots, n$. Using this and tailor expansion of the exponential function around zero, equation (1.15) reduces to

$$\begin{aligned}\Pr(M_n \le z) &\approx \exp\left(-\sum_{j=k}^{n} \alpha_{kj}(z) - \sum_{i=1}^{k-1} \alpha_{ii}(z)\right) \\ &\approx \exp\left(-\sum_{j=k}^{n} \alpha_{kj}(z)\right) \qquad (1.16)\end{aligned}$$

where $\alpha_{kj}(z) = \Pr(X_j \ge z | X_{j-1} \le z, \ldots, X_{j-k+1} \le z)$ for $k \ge 2$ and $\alpha_{kj}(z) = \Pr(X_j \ge z)$ for $k = 1$. The final step is justified since $\sum_{i=1}^{k-1} \alpha_{ii}(z)$ is negligible compared to $\sum_{j=k}^{n} \alpha_{kj}(z)$ for $k \ll n$, while the tailor expansion around zero is reasonable at the upper tail since for large z, $\alpha_{kj}(z)$ is close to zero.

Considering the Average Conditional Exceed Rate (ACER) as

$$\epsilon_k(z) = \frac{1}{n-k+1} \sum_{j=k}^{n} \alpha_{kj}(z). \qquad (1.17)$$

The ACER function can be estimated using

$$\hat{\epsilon}_k(z) = \frac{\sum_{j=k}^{n} \mathbf{1}(x_j \ge z, x_{j-1} \le z, \ldots, x_{j-k+1} \le z)}{\sum_{j=k}^{n} \mathbf{1}(x_{j-1} \le z, \ldots, x_{j-k+1} \le z)}. \qquad (1.18)$$

where $\mathbf{1}(\omega)$ is the indicator function for event ω. For nonstationary observations it is suggested using $n-k+1$ as an approximation for the denominator. The approximation can be justified since $\mathbf{1}(x_{j-1} \le z, \ldots, x_{j-k+1} \le z) \to 1$ in the upper tail where z is large.

It is assumed that the tail of the ACER function follows

$$\epsilon_k(z | a_k, b_k, c_k, q_k, \xi_k) = q_k \left[1 + \xi_k \left(a_k (z - b_k)^{c_k}\right)\right]^{-1/\xi_k}, \qquad (1.19)$$

where the parameters a_k, b_k, c_k, q_k and ξ_k are approximately constant in the upper tail for a certain k. The process of selecting k can be done by investigating the plot of $\hat{\epsilon}_k(z)$ against z for a verity of k, k is set to the smallest value for which increasing k makes negligible change to the tail. see figure REF TIL FIGUR PLUS SOME TEXT HER!!!!!!!!!!. The parameters can be estimated by minimizing the weighted square error

$$F(a,b,c,q,\xi) = \sum_{i=1}^{N} w_i \left[\log(\hat{\epsilon}_k(z_i)) - log(q) + \xi^{-1}\log\left(1 + a(z_i - b)^c\right)\right]^2, \tag{1.20}$$

using numerical methods. Selecting $z_1, \ldots, z_N$ is done by uniformly dividing the values from where regular tail behavior of $\hat{\epsilon}_k(z)$ starts to $\max_{1\le i\le n}(X_i)$ into N points. The weights w_i is calculated using

$$w_i = \left(\log\left[C_\alpha^+(z_i)\right] - \log\left[C_\alpha^-(z_i)\right]\right)^{-2}, \tag{1.21}$$

where $C_\alpha^+(z_i)$ and $C_\alpha^-(z_i)$ is the upper and lower $100 \cdot \alpha\%$ confidence interval values respectively for $\hat{\epsilon}_k(z_i)$. Organizing the observation into R similar realizations, like R years, the sample variance can be calculated as

$$\hat{s}_k(z_i)^2 = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\epsilon}_k^{(r)}(z_i) - \hat{\epsilon}_k(z_i)\right) \tag{1.22}$$

where $\hat{\epsilon}_k^{(r)}(z_i)$ is the estimated ACER function for the r realization at z_i. Hence a $100 \cdot \alpha\%$ confidence interval can be calculated using the student t-distribution

$$C_\alpha^\pm(z_i) = \hat{\epsilon}_k(z_i) \pm t_{(1-\alpha)/2,R-1}\frac{\hat{s}_k(z_i)}{\sqrt{R}} \tag{1.23}$$

where $t_{p,v}$ is defined as $\Pr(T > t_{p,v}) = p$ for the standardized t distribution with v degrees of freedom.

After parameter estimation, future prediction can be achieved using equation (1.19). Confidence intervals can be added to the ACER function prediction by estimating the parameters to the upper and lower confidence curve. Using $\epsilon_k(z_i|a,b,c,q,\xi) \pm t_{(1-\alpha)/2,R-1}\frac{\hat{s}_k(z_i)}{\sqrt{R}}$ instead of $\hat{\epsilon}_k(z_i)$ in equation (1.20), where $\epsilon_k(z_i|a,b,c,q,\xi)$ is given by equation (1.19), parameters for upper and lower confidence curves are estimated. These upper and lower confidence parameters can be used in equation (1.19) for ACER function prediction confidence intervals.

HUSK SKRIV OM PLOT FOR ACER TIL z I STEDET FOR η!!!!!!!!!!

1.1.3 Predicting Future Extreme With Confidence Interval

highest posterior density interval (Choosing the narrowest interval)

1.2 Bayesian Inference

For the traditional frequentist statistics, the parameters $\boldsymbol{\theta}$ are assumed fixed, while data $x_1, \ldots, x_n$ are random from the underlying distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$. Bayesian statistics, instead treats the parameters $\boldsymbol{\theta}$ with a probability distribution, where it is possible to make subjective believes about the distribution, independent of the data. These subjective beliefs is used to construct a prior distribution $f(\boldsymbol{\theta})$ based on experiences, information or physical knowledge of the situation analyzed.

The posterior distribution of the parameters dependent on the measured data becomes

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\theta}, \tag{1.24}$$

where Θ is the domain over all possible parameters, for which the integral is taken, and $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function. The likelihood function is constructed from the joint density function, which for independent data equals

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}). \tag{1.25}$$

The integral over parameters reduces to a constant, which makes

$$f(\boldsymbol{\theta}|\boldsymbol{x}) \sim c \cdot f(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta}), \tag{1.26}$$

where $c = 1/\int_\Theta f(\boldsymbol{\theta})f(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{d}\theta$ is the normalizing constant.

A conjugate prior is a prior which combined with the likelihood function construct a posterior distribution in the same family as the prior. Conjugate priors are often preferred because of the analytical luxury and computational simplicity.

Since Bayesian inference accounts for the distribution of parameters, equation (1.13) can be rewritten using $\Pr(X > u) = \alpha$

$$\Pr(X > z) = \alpha \left[1 + \xi \left(\frac{z-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \tag{1.27}$$

where α, ξ and σ are the unknown parameters. Since α is the probability of a point being larger than the threshold, α is independent of ξ and σ. For the independent parameters, development of the posterior distributions can be treated separately.

Starting with ξ and σ, by combining equation (1.25) and (1.10), the joint density function for the POT method becomes

$$\begin{aligned} f(\boldsymbol{y}|\xi,\sigma) &= \prod_{i=1}^{n} h(y_i|\xi,\sigma) \\ &= \sigma^{-n} \prod_{i=1}^{n} \left(1 + \frac{\xi y_i}{\sigma}\right)^{-\left(1+\frac{1}{\xi}\right)}, \end{aligned} \tag{1.28}$$

or by (1.11) for $\xi = 0$, $f(\boldsymbol{y}|\sigma) = \sigma^{-n} \exp\left\{-\sigma^{-1} \sum_{i=1}^{n} y_i\right\}$. here h is the probability density function of the GPD, while n is the numbers of threshold excess.

The obvious beginning for investigating priors is the conjugate priors, but unfortunately, there do not appear to be any conjugate priors for the joint GPD. This work, will not go into depth on how to select Bayesian priors for the GPD, but instead use the suggestion proposed by (Coles, 2001, p. 174). Note that there are potential improvements by deeper investigation of GPD or GEV priors, especially for priors developed for specific situations where there are physical knowledge or practical experiences about the parameters. Since $\sigma > 0$, the transformation $\phi = \log(\sigma)$ ensures σ to be valid without restriction on ϕ. The suggested priors are $f_\phi(\cdot)$ and $f_\xi(\cdot)$ to be normally distributed around zero with variance $v_\phi = 10^4$ and $v_\xi = 100$.

Considering the prior distribution of ϕ instead of σ, the change of variable for the joint density function becomes

$$\begin{aligned} f_{\boldsymbol{y}|\xi,\phi}(\boldsymbol{y}|\xi,\phi) &= \frac{f_{\boldsymbol{y},\xi,\phi}(\boldsymbol{y},\xi,\phi)}{f_{\xi,\phi}(\xi,\phi)} \\ &= \frac{f_{\boldsymbol{y},\xi,\sigma}\left(\boldsymbol{y},\xi,\exp(\phi)\right)\cdot\left|\frac{\mathrm{d}}{\mathrm{d}\phi}\exp(\phi)\right|}{f_{\xi,\sigma}\left(\xi,\exp(\phi)\right)\cdot\left|\frac{\mathrm{d}}{\mathrm{d}\phi}\exp(\phi)\right|} \\ &= f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi,\exp(\phi)\right), \end{aligned} \tag{1.29}$$

where $f_X(\cdot)$ indicates the probability distribution of X. The posterior distribution of the parameters can then be developed by the priors, (1.29) and (1.26)

$$f_{\xi,\phi|\boldsymbol{y}}(\xi,\phi|\boldsymbol{y}) \sim c\cdot f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi,\exp(\phi)\right) f_{\xi}(\xi) f_{\phi}(\phi), \tag{1.30}$$

where again c is the normalizing constant, $f_{\boldsymbol{y}|\xi,\sigma}$ is as in (1.28), $f_{\xi}(\xi) \sim N(0,100)$ and $f_{\phi}(\phi) \sim N(0,10^4)$.

The development of α posterior distribution, can be started with investigating priors. Since α simply equals $\Pr(X > u)$, the range is limited between 0 and 1. For simplicity the prior is set proportional to the uniform distribution on the interval $(0,1)$, $f(\alpha) \sim UNIF(0,1)$. It is noted that in reality the distribution of α is not flat. Low valued α is more probably then high, while the probability converges to zero for the endpoints. A well-tuned Beta distributed prior could account for this, and improve the result.

The joint density function can be created by the fact that α simply equals the probability of a random point exceeding the threshold. This can be expressed using the binominal distribution, k_i for the numbers of points exceeding the threshold and N_i for the total numbers of point, each for a given period i. For a total m numbers of periods, the posterior distribution equals

$$\begin{aligned} f(\alpha|k_1,\ldots,k_n,N_1,\ldots,N_n) &\sim f(\alpha)\cdot\prod_{i=1}^{m} f(k_i|N_i,\alpha) \\ &= \prod_{i=1}^{m}\binom{N_i}{k_i}\alpha^{k_i}(1-\alpha)^{N_i-k_i} \\ &\sim \alpha^{\sum_{i=1}^{m}k_i}(1-\alpha)^{\sum_{i=1}^{m}N_i-\sum_{i=1}^{m}k_i}. \end{aligned} \tag{1.31}$$

The resulting distribution is independent of period selection, the notation k and N can be used for the total numbers of exceedance and the total numbers of measurements respectively. It is noted that the distribution is proportional to the Beta distribution. Since there only exist one normalizing constant which fulfills the requirements for a probability distribution, the posterior distribution is not only proportional, but equal to the Beta distribution. Rewriting equation (1.31) gives

$$f(\alpha|k,N) \sim BETA(k+1, N-k+1). \tag{1.32}$$

1.3 Markov Chain Monte Carlo Method

Markov Chain Monte Carlo (MCMC) is a powerful iterative method used to sample from probability distributions, which analytically or through other simulation methods, could be difficult and impracticable to sample from. The algorithm is constructed by converging the desired probability distribution to an irreducible and aperiodic Markov Chain with limiting distribution equal the target distribution. Independent of starting position the Markov Chain will then converge towards the desired probability distribution in the limit as the numbers of iterations goes to infinity. The first numbers of realizations from the Markov Chain until converging is called the burn-in period, and is discarded for further analyses. More information about burn-in can be found in (Givens and Hoeting, 2013, p.220). The remaining realizations, approximately follows the desired probability distribution, where the accuracy increase as the numbers of realizations increase. Monte Carlo method can then be used to calculate the quantity of interest like mean, expected value, future prediction, credible interval etc. For more in depth description of the MCMC method see the books of Gamerman and Lopes (2006) and (Givens and Hoeting, 2013, Chapter 7,8).

1.3.1 Gibbs Sampling

In situations where it is difficult to sample from the joint distribution, but applicable from the conditional distribution, Gibbs sampler is preferable. The theory behind Gibbs sampling was first proposed in Geman and Geman (1984). The principle of Gibbs sampling is to construct the Markov Chain, by repeatedly sample each parameter with the rest of the parameters as the condition. The Gibbs sampler start with an initial guess $\boldsymbol{X}^0 = [X_1^0, \cdots, X_n^0]$ for the parameters $\boldsymbol{X}$, then iteratively update each as follows

$$\begin{aligned}
X_1^{t+1}|\cdot &\sim f(X_1|X_2^t, \dots, X_n^t), \\
X_2^{t+1}|\cdot &\sim f(X_2|X_1^{t+1}, X_3^t, \dots, X_n^t), \\
&\vdots \\
X_n^{t+1}|\cdot &\sim f(X_n|X_1^{t+1}, \dots, X_{n-1}^{t+1}).
\end{aligned} \tag{1.33}$$

where t is the iteration number, f is probability function of the parameter and $|\cdot$ symbolize that the function is conditional on the rest and recent parameters. In some cases, it could be beneficial to sample some of the parameters in blocks, such as $(X_k, X_{k+1})|\cdot$ where $1 \le k \le n$. This form of Gibbs sampling is called blocking. The iterative process is repeated until enough realizations are generated for sufficient accuracy. More on Gibbs sampling can be found in (Gamerman and Lopes, 2006, p. 141) and (Givens and Hoeting, 2013, p. 209)

1.3.2 Metropolis–Hastings Algorithm

The Metropolis-Hastings algorithm was first proposed by Metropolis et al. (1953), and is another method for constructing a suitable Markov Chain. The algorithm is preferable for situations where a proportional distribution is simple to evaluate, while the target probability distribution is difficult. Bayesian inference (see chapter 1.2), often result in a distribution where the normalizing constant cannot analytically be calculated. While possible numerically, the normalizing constant often becomes computationally hard, which make it impracticable for iterative simulations. Using Metropolis–Hastings algorithm on a proportional distribution whiteout normalizing constant, results in samples from the target distributions.

The Metropolis–Hastings algorithm start with an initial guess for the parameters. For each iteration a new parameters $\boldsymbol{X}^*$ are suggested from a proposal distribution $g(\boldsymbol{X}^*|\boldsymbol{X}^t)$, given the last accepted parameter $\boldsymbol{X}^t$. The new parameter is then evaluated against the last accepted by

$$R(\boldsymbol{X}^*, \boldsymbol{X}^t) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{X}^t|\boldsymbol{X}^*)}{f(\boldsymbol{X}^t)g(\boldsymbol{X}^*|\boldsymbol{X}^t)} \tag{1.34}$$

where $f(x)$ is the target distribution, or a distribution proportional to the target distribution. The parameter $\boldsymbol{X}^{t+1}$ takes value $\boldsymbol{X}^*$ with probability $\min\{1, R(\boldsymbol{X}^*, \boldsymbol{X}^t)\}$, if rejected, $\boldsymbol{X}^{t+1} = \boldsymbol{X}^t$ instead. The reason the target distribution normalizing constant is irrelevant, is because they are both canceled out in $f(\boldsymbol{X}^*)/f(\boldsymbol{X}^t)$.

A common proposal distribution is the random walk. The new parameters are generated from the last accepted realization with additional variance, $\boldsymbol{X}^* = \boldsymbol{X}^t + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ follows a chosen probability distribution. Symmetric Proposals implies that $g(\boldsymbol{X}^t|\boldsymbol{X}^*) = g(\boldsymbol{X}^*|\boldsymbol{X}^t)$, this is referred to as Metropolis algorithm.

The Metropolis–Hastings algorithm could in situations be necessary for some of the steps in the Gibbs sampler, equation (1.33). Such a combination of Metropolis-Hastings algorithm and Gibbs sampler is referred to as a Hybrid Gibbs sampler, and was first introduced by Müller (1991). For more information about the Metropolis-Hastings algorithm and Hybrid Gibbs sampler see (Givens and Hoeting, 2013, p. 202), (Gamerman and Lopes, 2006, p. 191) and (Givens and Hoeting, 2013, p. 216), cite[p. 205]MCMC respectively.

1.3.3 Effective Sample Size

The realizations of the simulated Markov Chain will often be correlated, and dependent on the future and past iterations. The correlation implies that the information hold by the realizations is actually less than the numbers of realizations. The effective sample size gives a method of calculating the theoretical size of an equally informative i.i.d., realization set. The effective sample size is estimated as

$$L_{eff} = \frac{L}{1 + 2\sum_{k=1}^{K} \hat{\rho}(k)}, \tag{1.35}$$

where L is the sample size of the simulated realizations, $\hat{\rho}(k)$ is the estimated k step autocorrelation between realizations and K is chosen as the first k where $\hat{\rho}(k) < 0.1$. The effective sample size is a quantification of the information hold by the simulated realization set.

1.3.4 Adaptive Metropolis Algorithm

A challenge with constructing a MCMC is to ensure that the series converge to the stationary target distribution relatively quickly, and that the samples gives points in the whole range of the target distribution, this is referred to as good mixing.

If a large percentage of Metropolis-Hastings proposal $\boldsymbol{X}^*$ is accepted, the proposal distribution is too narrow. High acceptance rate will delay convergence, and cause higher correlation between points. The result is poor mixing and a decrease in effective sample size. On the other hand, if only a small percentage of the proposals is accepted, the proposal distribution is to wide. Low acceptance rate will also increase correlation, which gives poor mixing and decreased effective sample size. A large number of generated realizations by the Markov chain will be equal, which will harm future Monte Carlo simulation.

To maximize the effective sample size and ensure good mixing, the acceptance rate should be somewhere in between. For a Metropolis-Hastings algorithm, Gelman et al. (1996) suggested a 44% acceptance rate for single dimensional normal target distribution and 23.4% for high dimensional multivariate normal target distribution. Commonly the user would run the Metropolis-Hastings algorithm, calculate acceptance rate, tune variance and then rerun the process until sufficient acceptance rate is achieved.

For this work, MCMC simulation will be used for a large number of different situations, and it would become extremely time-consuming to tune each variance. The inconvenient can be handled by using an Adaptive Markov Chain Monte Carlo (AMCMC) which adapt the MCMC algorithm while running. This is achievable using a normal random walk proposal where the next suggested realization $\boldsymbol{X}^* \sim N(\boldsymbol{X}^t, \lambda\boldsymbol{\Sigma}^t)$. Between iterations $\boldsymbol{\Sigma}^t$ is adjusted to improving mixing and efficient sample size. The acceptance rate is set by λ, and with a p dimensional multivariate normal target distribution, it has been shown that a constant $\lambda = 2.38^2/p$ is optimal when $\boldsymbol{\Sigma}$ equals the real variance of the target distribution, Gelman et al. (1996). The adaptive Metropolis algorithm is not constraint to the normal target distributions, but the suggested λ seems like a good starting value. The ability of an adjustable λ between future iteration seems beneficial, because of the unknown target distribution and the following acceptance rate. The additional adaptive parameter $\boldsymbol{\mu}^t$ is necessary since the covariance is proportional to $\boldsymbol{\mu}$. The initial guess is chosen as $\boldsymbol{\mu}^0 = \mathbf{0}$ and $\boldsymbol{\Sigma}^0 = \mathbf{I}$. The normal random walk proposal distribution is symmetric, which result in an adaptive Metropolis algorithm, where (1.34) is reduced to

$$R(\boldsymbol{X}^*, \boldsymbol{X}^t) = \frac{f(\boldsymbol{X}^*)}{f(\boldsymbol{X}^t)}. \tag{1.36}$$

For each iteration $\boldsymbol{\mu}^{t+1}$ and $\boldsymbol{\Sigma}^{t+1}$ is updated as follows

$$\boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^{t} + \gamma^{t+1}(\boldsymbol{X}^{t+1} - \boldsymbol{\mu}^{t}) \tag{1.37}$$

$$\boldsymbol{\Sigma}^{t+1} = \boldsymbol{\Sigma}^{t} + \gamma^{t+1}\left[(\boldsymbol{X}^{t+1} - \boldsymbol{\mu}^{t})(\boldsymbol{X}^{t+1} - \boldsymbol{\mu}^{t})^{T} - \boldsymbol{\Sigma}^{t}\right], \tag{1.38}$$

Where γ is a decreasing parameters which provide the Markov chain property described in the beginning of chapter 1.3. The details of γ^t, to ensure an irreducible and aperiodic Markov chain can be found in Roberts and Rosenthal (2007) and Atchadé et al. (2011) , where it is noted that $\lim_{t->\infty}\gamma^t = 0$ while the summation, not necessary is bounded $\sum_{t=1}^{\infty}\gamma = \infty$. Repeated trails concluded that $\gamma^t = \cdots$ HUSK AA FYLL INN HER!!!!!!!!!! was sufficient choice.

As described above an adaptive λ^t could be beneficial. By using

$$\log(\lambda^{t+1}) = \log(\lambda^{t}) + \gamma^{t+1}\left(R(\boldsymbol{X}^{*}, \boldsymbol{X}^{t}) - a\right), \tag{1.39}$$

the series acceptance rate will converge towards a (Givens and Hoeting, 2013, p. 248).

A more detailed description of the adaptive metropolis algorithm can be found in (Givens and Hoeting, 2013, p. 247). CHECK BOLDSYMBOL FOR VECTOR AND MATRIX!!

1.3.5 Applying Markov Chain Monte Carlo to Peak Over Threshold

From chapter 1.2, the two equation (1.30) and (1.32) construct the basis for the blocking Gibbs sampler

$$\xi^{t+1}, \phi^{t+1}|\cdot \sim c \cdot f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi, \exp(\phi)\right) f_{\xi}(\xi) f_{\phi}(\phi)$$
$$\alpha^{t+1}|\cdot \sim BETA(k+1, N-k+1),$$

where the parameters and functions are described above in chapter 1.2. After the Markov chain sampling is complete, the transformation $\sigma = \exp(\phi)$ ensure correct parameter for the Monte Carlo simulation's. Sampling form α is straight forward since it is simply realizations of the Beta distribution, while ξ, σ is more complex, and cannot directly be sampled. The challenge of calculating the computationally heavy c for each iteration favors the implementation of the

Metropolish-Hastings algorithm.

The algorithm independency of user dependent tuning and improved convergence speed makes the adaptive Metropolis-Hasting algorithm, described in chapter 1.3.4, favorable for ξ, σ. The posterior distribution often results in extremely small values, which in some cases could get disrupted by the violation of the smallest floating number for the software. To account for this, the logarithm of equation (1.36) is used. The resulting log Metropolis ratio

$$\begin{aligned}\ln\left[R(\boldsymbol{X}^*, \boldsymbol{X}^t)\right] &= \ln\left[f_{\xi,\phi|\boldsymbol{y}}(\xi^*, \phi^*|\boldsymbol{y})\right] - \ln\left[f_{\xi,\phi|\boldsymbol{y}}(\xi^t, \phi^t|\boldsymbol{y})\right] \\ &= \ln\left[f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi^*, \exp(\phi^*)\right)\right] - \ln\left[f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi^t, \exp(\phi^t)\right)\right] + \\ &\quad \ln\left[f_\xi(\xi^*)\right] + \ln\left[f_\phi(\phi^*)\right] - \ln\left[f_\xi(\xi^t)\right] - \ln\left[f_\phi(\phi^t)\right],\end{aligned} \tag{1.40}$$

where equation (1.28) gives,

$$\ln\left[f_{\boldsymbol{y}|\xi,\sigma}\left(\boldsymbol{y}|\xi, \exp(\phi)\right)\right] = \begin{cases} -n\phi - \exp(-\phi)\sum_{i=1}^n y_i, & \xi = 0, \\ -n\phi - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^n \ln\left(1 + \xi\exp(-\phi)y_i\right), & \text{else}, \end{cases} \tag{1.41}$$

and after inserting the priors mean and variance from chapter 1.2, the remaining parts reduces to

$$\ln\left[f_\xi(\xi^*)\right] + \ln\left[f_\phi(\phi^*)\right] - \ln\left[f_\xi(\xi^t)\right] - \ln\left[f_\phi(\phi^t)\right] = -\frac{(\xi^*)^2 - (\xi^t)^2}{200} - \frac{(\phi^*)^2 - (\phi^t)^2}{2\cdot 10^4}. \tag{1.42}$$

For the equations, t indicate the parameter iteration number, * indicate the proposed parameter value to be evaluated, ξ and σ are GPD parameters where $\phi = \log(\sigma)$ and $\boldsymbol{y}$ is a vector containing the observed data of size n.

The remaining construction of the AMCMC simmulator is as described above in chapter 1.3.4. The result is a hybrid Gibbs AMCMC simulator, for the POT method. The Markov chain is valid since both Gibbs steps are irreducible and aperiodic. Irreducible because each sampler within their restricted range can sample any realization with probability larger than zero, from any state. Aperiodic since both Gibbs steps can return to their state in a single iteration, with probability larger than zero.

1.3.6 Multivariate Random Normal Generator

Most of the coding for this work was done in R. Since the MCMC accuracy increase with the numbers of iteration generated, parts of the AMCMC algorithm was coded in C++, through the Rcpp package by Eddelbuettel et al. (2016), for speed optimization. The AMCMC for ξ and ϕ uses a bivariate random normal generator, but this is not natively supported in C++. Since no additional packages tested was satisfactory for the purpose, the bivariate random normal generator was constructed.

The Box-Muller transformation, see Box and Muller (1958), states that for two independent random variables $U_1, U_2 \sim UNIF(0,1)$, the transformation

$$z_1 = \sqrt{-2\ln(U_1)}\, \cos(2\pi U_2)$$
$$z_2 = \sqrt{-2\ln(U_1)}\, \sin(2\pi U_2)$$

results in Z_1 and Z_2 to be independent and standard normally distributed. Combined in a vector $\boldsymbol{z} = [Z_1, Z_2]^T$ where T is the transpose, the bivariate random normal $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be generated by

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{z}, \tag{1.43}$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$. For this work $\boldsymbol{A}$ is chosen as the Cholesky decomposition of $\boldsymbol{\Sigma}$.

STATE SOMEWHERE THAT log IS THE NATURAL LOGRATIM?

1.4 Value at Risk

mer bla bla

1.5 Evaluating Forecasts

blablabla likelihood ratio, Kupiec (1995), Christoffersen (1998).

And scoring table (how to).

1.6 ARCH/GARCH

is this needed????

Chapter 2

Data

hei

2.1 Synthetic Data

blabla

2.1.1 Pareto

sim sim

2.1.2 Commodity Generated Data

sam sam

2.2 Commodity Data

Adjusted for the roll-over returns. (simply done by deleting the return at the roll-over date).(roll-over, en kontrakt til neste eks sommer til hoest, kan gi ekstreme returns som gir forvrengende resultater). blabla

Chapter 3

Summary and Recommendations for Further Work

Better prior MCMC (Improved prior)=> higher effective sample size.

GEV instead of POT! (since AR-GARCH ~ i.i.d).

Improved ACER program. ACER continious CI so qq-plot is possible (given a 1% event, where on the extreme prediction confidence level is that)

Automatically selecting u for POT MCMC each turn.

REMEMBER IN THEORY, WHY ADAPTIVE MCMC!! MAX EFFECTIVE SAMPLE SIZE, AND MORE IMPORTANTLY AUTOMATIC!!!!! MINIMAL HUMA INTERACTION.

3.1 Summary and Conclusions

Here, you present a brief summary of your work and list the main results you have got. You should give comments to each of the objectives in Chapter 1 and state whether or not you have met the objective. If you have not met the objective, you should explain why (e.g., data not available, too difficult).

This section is similar to the Summary and Conclusions in the beginning of your report, but more detailed—referring to the the various sections in the report.

3.2 Discussion

Here, you may discuss your findings, their strengths and limitations.

3.3 Recommendations for Further Work

You should give recommendations to possible extensions to your work. The recommendations should be as specific as possible, preferably with an objective and an indication of a possible approach.

The recommendations may be classified as:

- Short-term
- Medium-term
- Long-term

Appendix A

Additional Information

This is an example of an Appendix. You can write an Appendix in the same way as a chapter, with sections, subsections, and so on.

A.1 Introduction

A.1.1 R

```
#Dette er en melding
string<-"Heisann du"
a=5
for(i in 1:a){
  print(i)
}
bla
blla
```

A.1.2 C++

```
#include<stdio.h>
#include<iostream>
// A comment
int main(void)
{
    for(int i=0;i<5;i++){
        cout<<i;
    }
    printf("Hello World\n");
    return 0;
}
```

Bibliography

Atchadé, Y., Fort, G., Moulines, E., and Priouret, P. (2011). Adaptive markove chain monte carlo: Theory and methods. In *Baysian Time Series Models*. Cambrudge Univeristy Press.

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*.

Christoffersen, P. F. (1998). Evaluating interval forecasts. In *International Economic Review*, volume 39, pages 841–862. Economics Department of the University of Pennsylvania.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Bates, D., and Chambers, J. (2016). *Rcpp: Seamless R and C++ Integration*.

Gamerman, D. and Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Champman and Hall/CRC, 2nd edition.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient metropolis jumping rules. In *Bayesian Statistics 5*.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Givens, G. H. and Hoeting, J. A. (2013). *Computational Statistics*. John and Sons, 2nd edition.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*.

Müller, P. (1991). A generic approach to posterior integration and gibbs sampling. Technical report, Department of Statistics, Purdue University.

Næss, A. and Gaidai, O. (2009). Estimation of extreme value from sampled time series. In *Structural Safety*, volume 31, pages 325–334.

Næss, A., Gaidai, O., and Karpa, O. (2013). Estimation of extreme value by the average conditional exceedance rate method. *Journal of Probability and Statistics*.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of Applied Probability*.